6S-07

URL 文字列の特徴量から ユーザーへの攻撃を検出する機械学習モデル

田原 舜大 † 秦野 亮 ‡ 西山 裕之 §
東京理科大学 理工学部 † 東京理科大学 理工学部 ‡ 東京理科大学 理工学部 §

1 序論

近年，サイバー攻撃は増加傾向にあり，攻撃手段の 1 つとして URL(Uniform Resource Locator) が用いられる．その中のフィッシング，Web サイト改竄，マルウェア，スパム，XSS(Cross Site Scripting) は URL にアクセスしたユーザーに直接的に被害を及ぼす攻撃手法である．そのような URL の検出にブラックリストが用いられていたが，新たな種類の URL が日々生み出されており，この手法では新たな脅威への対応で漏れが生じてしまう．そのような漏れを減らすために本研究では，URL 文字列から作成した特徴量からユーザーへの攻撃を検出する機械学習モデルを構築する．

2 関連研究

Banerjee ら [1] は，URL 及び JavaScript から作成した特徴量から機械学習を用いて XSS の検知を行った．しかし，JavaScript に関するデータを取得するためには URL にアクセスする必要があり，攻撃を受けることが前提となっている．Jagdale ら [2] は，URL 文字列から作成した特徴量からアンサンブル学習を用いてフィッシング検出を行った．特徴量には，URL の長さ，“-” が存在するか否か，“.” の個数などが用いられた．しかし，同研究はフィッシング検出に焦点を置いた特徴量作成をしているため，別の攻撃手法への有効性が担保されていない．そこで本研究では，URL 文字列のみから作成した特徴量で，ユーザーに被害を及ぼすリスクのある URL をより汎用的に検出する機械学習モデルを構築する．

3 提案手法

3.1 データセット

本研究では，ISCX-URL2016 データセット [3] と XSS データセット [4] を利用する．上記のデータセットから URL 文字列データと URL の種類を合計 193,033 件抽出する．そして，取得したデータの URL の種類からラベル付けをする．URL の種類が何かしらの攻撃手法 (フィッシング，Web サイト改竄，マルウェア，スパム，XSS) を示している場合は「有害」，それ以外を「無害」とラベル付けする．その結果，「有害」ラベルは 144,236 件，「無害」ラベルは 48,797 件となった．次に，作成したデータセットを学習データとテストデータに 7:3 の割合で分割し，それぞれのデータに対して特徴量作成を行う．

3.2 特徴量作成

本研究では，Jagdale ら [2] が使用した特徴量に加えて，次の系統の特徴量を新たに作成した．

- 文字列の長さ
- uni-gram の one-hot エンコーディング
- uni-gram の count エンコーディング
- パーセントエンコーディングされている特殊文字の数

A machine learning model for detecting attacks on users based on features of URL string
† Toshimasa Tahara, Tokyo University of Science
‡ Ryo Hatano, Tokyo University of Science
§ Hiroyuki Nishiyama, Tokyo University of Science

表 2　誤検出率 (%)

	Web サイト改竄	スパム	XSS	マルウェア	フィッシング	無害
Jagdale ら [2]	2.92	0.17	3.39	4.84	1.64	4.32
本研究	0.66	0.06	0.72	0.49	1.10	2.27

uni-gram の one-hot エンコーディングとは，URL 内のある 1 文字 (uni-gram) の有無を 0,1 で表現するものであり，count エンコーディングとは，ある 1 文字の出現回数を表現する．また，パーセントエンコーディングとは，URL 内の特殊文字 (“:” など) をエンコードする仕組み (一種のエスケープ) であり，他の系統には当てはまらない特徴であるため，別の系統として区別している．

3.3 機械学習モデル

本研究では，Jagdale ら [2] と同様に勾配ブースティングと決定木のスタッキングを分類器として用いる．また，評価指標には，正解率，再現率，適合率，F1 値を使用し，URL の種類ごとの誤検出率も算出した．

4 実験

実験では，Jagdale ら [2] が採用した特徴量のみのモデルと，それに本研究で提案した特徴量を追加したモデルの精度を比較した．各評価指標の値を表 1 に示す．本モデルの再現率は，98.95% であり Jagdale ら [2] よりも 2.17 ポイント優れていることがわかる．再現率は，有害な URL 検出の網羅性を意味しているため，本モデルはより検出漏れの少ないモデルであるといえる．また，適合率は，無害な URL を有害な URL と検知した数の少なさを意味しており，これが低いと攻撃性のないサイトを危険だと認識してユーザーの利便性を損なう．表 1 より，本モデルの方が Jagdale ら [2] よりも 0.72 ポイント優れていることがわかるため，本モデルはユーザーの利便性を損なわない．そして，再現率と適合率のバランスを示す F1 値が 99.30% であることからも，本研究のモデルの方が Jagdale ら [2] のモデルよりも攻撃検出において総合的に優れた手法であるといえる．

表 1　各評価指標の値 (%)

	正解率	再現率	適合率	F1 値
Jagdale ら [2]	96.82	97.20	98.52	97.85
本研究	98.95	99.37	99.24	99.30

表 2 は，URL の種類ごとの誤検出率を表しており，本モデルの方がフィッシング含むいずれの攻撃手法の誤検出率が低くなっていることがわかる．これより，本モデルの方がフィッシングに限らない，より多くの攻撃手法の検出に対して有効であるといえる．

5 結論

本研究では，ユーザーに直接的な被害を及ぼす恐れのある有害な URL を検出する機械学習モデルを構築した．実験では，Jagdale ら [2] のモデルとの比較を行い，本研究のモデルの有効性を示した．今後の展望として，データ数の増加や SHAP(SHapley Additive exPlanations) を用いた特徴量選定などが挙げられる．

参考文献

[1] R. Banerjee, A. Baksi, N. Singh and S. K. Bishnu: Detection of XSS in web applications using Machine Learning Classifiers. 2020 4th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), Kolkata, India, 2020, pp. 1-5.
[2] N. Jagdale and P. Chavan: Hybrid Ensemble Machine Learning Approach for URL Phishing Detection. 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 2022, pp. 1-8
[3] URL dataset (ISCX-URL2016) https://www.unb.ca/cic/datasets/url-2016.html
[4] fmereani/Cross-Site-Scripting-XSS https://github.com/fmereani/Cross-Site-Scripting-XSS